Chapter 4: More on Two-Variable Data

4.2: Cautions about Correlation & Regression

* **Correlation and Regression**
  - Describe only linear relationships
  - Are not resistant
    - One influential observation or incorrectly entered data point can greatly change these measures.
* **Extrapolation**
  - Use of a regression line for prediction far outside the domain
  - Such predictions are not generally accurate

2

Cautions about Correlation & Regression *cont.*

- **Lurking Variable**
  - Variable not among the explanatory or response variables
  - May influence the interpretation of relationships among those variables.
- **Confounding Variable**
  - Two variables are confounded when their effects on a response variable cannot be distinguished from each other
  - May be either explanatory variables or lurking variables

3

Cautions about Correlation & Regression *cont.*

- **Studies Using Averaged Data**
  - Resist *(don't)* applying results to individuals
    - Correlations usually too high when applied to individuals
- **Causation (cause & effect)**
  - Even a very strong association between two variables is *not* by itself good evidence that there is a cause-and-effect link
    - High correlation *DOES NOT* imply causation
  - Establishing Causation
    - Conduct a carefully designed experiment
    - Control effects of possible lurking variables

4

Section 4.2 Complete

* Homework: #'s 27, 28, 33, 36
* Any questions on pg. 25-28 in additional notes packet

5

Section 4.3: Relations in Categorical Data

* **Categorical Variables**
  - Use counts or percentages that fall into various categories
  - Organized into two-way tables
    - Two-way tables describe two categorical variables
    - Rows make up one variable; columns make up the other

		One Variable with its categories				
Second variable		Cat 1	Cat 2	Cat 3	Cat 4	Total
	Cat 1	Conditional Distributions will be created by taking these values and dividing by either one of the Marginal Distributions				Marginal Distributions
	Cat 2					
	Total	Marginal Distributions				Total

6

4.53 SMOKING BY STUDENTS AND THEIR PARENTS Here are data from eight high schools on smoking among students and among their parents:[28]

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	1168	1823	1380
Student smokes	188	416	400
	1356	2239	1780
	25.2%	41.7%	33.1%

(a) How many students do these data describe? 5375

(b) What percent of these students smoke? $\frac{1004}{5375} = 18.7\%$

(c) Give the marginal distribution of parents' smoking behavior, both in counts and in percents.

7

4.53 SMOKING BY STUDENTS AND THEIR PARENTS Here are data from eight high schools on smoking among students and among their parents:[28]

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	26.7%	41.7%	31.6%
Student smokes	188	416	400
	1356	2239	1780
	25.2%	41.7%	33.1%

(a) How many students do these data describe? 5375

(b) What percent of these students smoke? $\frac{1004}{5375} = 18.7\%$

(c) Give the marginal distribution of parents' smoking behavior, both in counts and in percents.

(d) Give the conditional distribution of parents' smoking behavior given the student does not smoke.

$\frac{1168}{4371} = 26.7\%$ $\frac{1823}{4371} = 41.7\%$ $\frac{1380}{4371} = 31.6\%$

8

4.54 PYTHON EGGS How is the hatching of water python eggs influenced by the temperature of the snake's nest? Researchers assigned newly laid eggs to one of three temperatures: hot, neutral, or cold. Hot duplicates the extra warmth provided by the mother python, and cold duplicates the absence of the mother. Here are the data on the number of eggs and the number that hatched:[29]

	Cold	Neutral	Hot
Number of eggs	27	56	104
Number hatched	16	38	75

(a) Make a two-way table of temperature by outcome (hatched or not).

(b) Calculate the percent of eggs in each group that hatched. The researchers anticipated that eggs would not hatch in cold water. Do the data support that anticipation?

	Cold	Neutral	Hot
Hatched	59%	68%	72%
Did not hatch	41%	32%	28%

9

Simpson's Paradox:

* Refers to the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group.
  - The principle behind this reversal is that of weighted averages. Using a higher weight in one category for one group while the second group uses a second category to be its higher weight.

10

An Example of Simpson's Paradox

* Upper Wabash Tech has two professional schools, business and law. Here are two-way tables of applicants to both schools, categorized by gender and admission decision.

Business		
	Admit	Deny
Male	480	120
Female	180	20

Law		
	Admit	Deny
Male	10	90
Female	100	200

A. Combine the data to make a single two-way table of gender by admission decision.

11

Cont.

Business			
	Admit	Deny	Percent
Male	480	120	80%
Female	180	20	90%

Law			
	Admit	Deny	Percent
Male	10	90	10%
Female	100	200	34%

Combined			
	Admit	Deny	Percent
Male	490	210	70%
Female	280	220	56%

B. What percent of males are admitted? What percent of females are admitted? Which percent is higher?

C. Now compute the percent males and females admitted by each of the separate schools. Which percents are higher?

12

Cont.

D. Why is this occurring? Explain how it can happen that Wabash appears to favor males when each school individually favors females.

* Ans: The business school accepts a higher percentage of students all together than the law school does. So when 86% of the males applied to the business school while only 40% of the females applied to the business school, more males got into Wabash university all together. This gives the appearance that Wabash favors males. This reversal of the direction of an association when data from several groups are combined to form a single group is know as Simpson's Paradox.

13

Section 4.3 Complete

* Homework: #'s 55, 59, 61, 62,

* Any questions on pg. 29-32 in additional notes packet

14

Chapter Review

In Chapter 3, we learned how to analyze two-variable data that show a linear pattern. We learned about positive and negative associations and how to measure the strength of association between two variables. We also developed a procedure for constructing a model (the least-squares regression line) that captures the trend of the data. This LSRL is useful for prediction purposes. A recurring theme is that data analysis begins with graphs and then adds numerical summaries of specific aspects of the data.

15

In this chapter we learned how to construct mathematical models for data that fit a curve, such as an exponential function or a power function. We also learned that although correlation and regression are powerful tools for understanding two-variable data when both variables are quantitative, both correlation and regression have their limitations. In particular, we are cautioned that a strong observed association between two variables may exist without a cause-and-effect link between them. If both variables are categorical, there is no satisfactory graph for displaying the data, although bar graphs can be helpful. We describe the relationship by comparing percents.

Here is a review list of the most important skills you should have gained from studying this chapter.

16

A. MODELING NONLINEAR DATA

1. Recognize that when a variable is multiplied by a fixed number greater than 1 in each equal time period, exponential growth results; when the ratio is a positive number less than 1, it's called exponential decay.

2. Recognize that when one variable is proportional to a power of a second variable, the result is a power function.

3. In the case of both exponential growth and power function, perform a logarithmic transformation and obtain points that lie in a linear pattern. Then use least-squares regression on the transformed points. An inverse transformation then produces a curve that is a model for the original points.

4. Know that deviations from the overall pattern are most easily examined by fitting a line to the transformed points and plotting the residuals from this line against the explanatory variable (or fitted values).

17

B. INTERPRETING CORRELATION AND REGRESSION

1. Understand that both r and the least-squares regression line can be strongly influenced by a few extreme observations.

2. Recognize possible lurking variables that may explain the observed association between two variables x and y.

3. Understand that even a strong correlation does not mean that there is a cause-and-effect relationship between x and y.

18

C. RELATIONS IN CATEGORICAL DATA

1. From a two-way table of counts, find the marginal distributions of both variables by obtaining the row sums and column sums.
2. Express any distribution in percents by dividing the category counts by their total.
3. Describe the relationship between two categorical variables by computing and comparing percents. Often this involves comparing the conditional distributions of one variable for the different categories of the other variable.
4. Recognize Simpson's paradox and be able to explain it.

19

Chapter 4 Complete

* Homework: #'s 79, 81, 83
* Any questions on pg. 33-36 in additional notes packet

20